PRIMAVERA Data Management and Analysis using JASMIN

Jon Seddon[1], Ag Stephens[2], Malcolm Roberts[1]

[1] Met Office, UK. [2] Science and Technology Facilities Council (STFC), UK.

PRIMAVERA is a European Union Horizon 2020 funded project that aims “to develop a new generation of advanced and well-evaluated high-resolution global climate models, capable of simulating and predicting regional climate with unprecedented fidelity, for the benefit of governments, business and society in general”.

PRIMAVERA includes the running of seven climate models with common initial conditions and common forcings at low and high-resolutions to generate a multi-model ensemble of climate simulations. The volume of data from these simulations is estimated at 2.5 petabytes. The simulations will be run on High Performance Computers (HPCs) throughout Europe. The project required a central location where data from these simulations could be brought together and analysed. JASMIN's fast disk storage, tape archive, extensive compute resources and high-bandwidth connections to Europe makes it the ideal platform for PRIMAVERA's data management and analysis requirements. JASMIN allows the 100 PRIMAVERA scientists from across Europe to collaborate together on this multi-model ensemble of high-resolution climate data.

The Data Challenge

The PRIMAVERA stream 1 simulations will produce 2.5 petabytes of high resolution model data. The project has 440 terabytes of group workspace storage available to it. Figure 1 shows the workflow of data in PRIMAVERA. Data is transferred across the Internet from the HPCs throughout Europe to JASMIN group workspaces using JASMIN’s high-performance data transfer service.

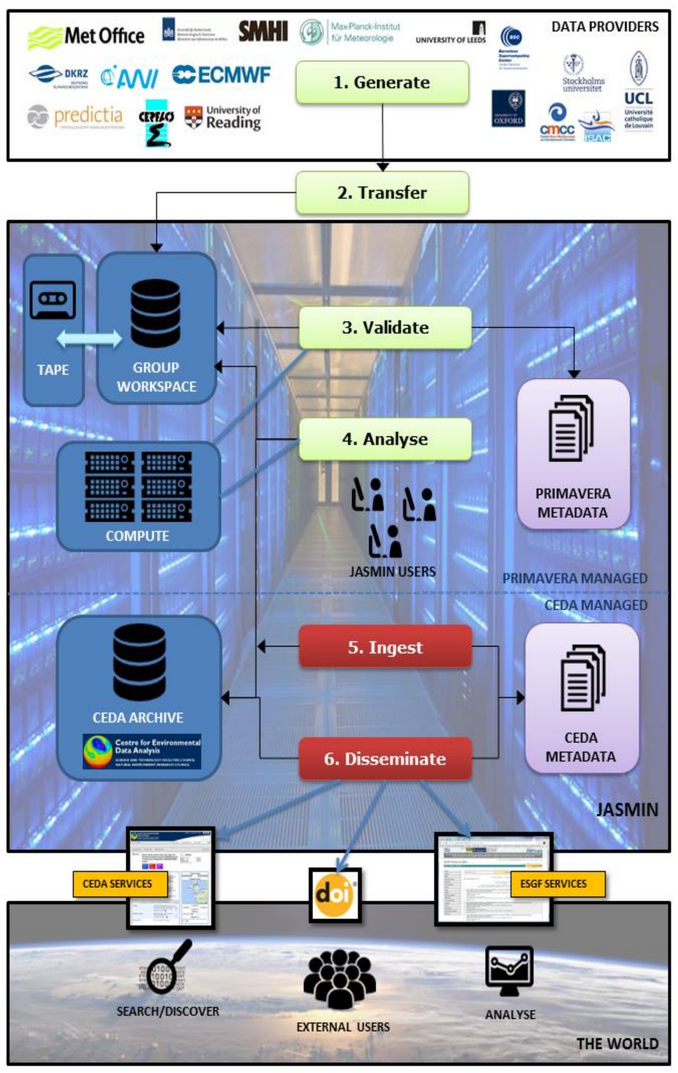


Figure 1. The workflow devised.

On arrival at JASMIN, data is validated and the metadata recorded in the database. The files are then written to elastic tape, making disk space available for further chunks of data to be transferred to JASMIN.

Users are now restoring the data from tape that they require for their work. Additionally, the data is being made available to the wider community through CEDA’s Earth System Grid Federation node.

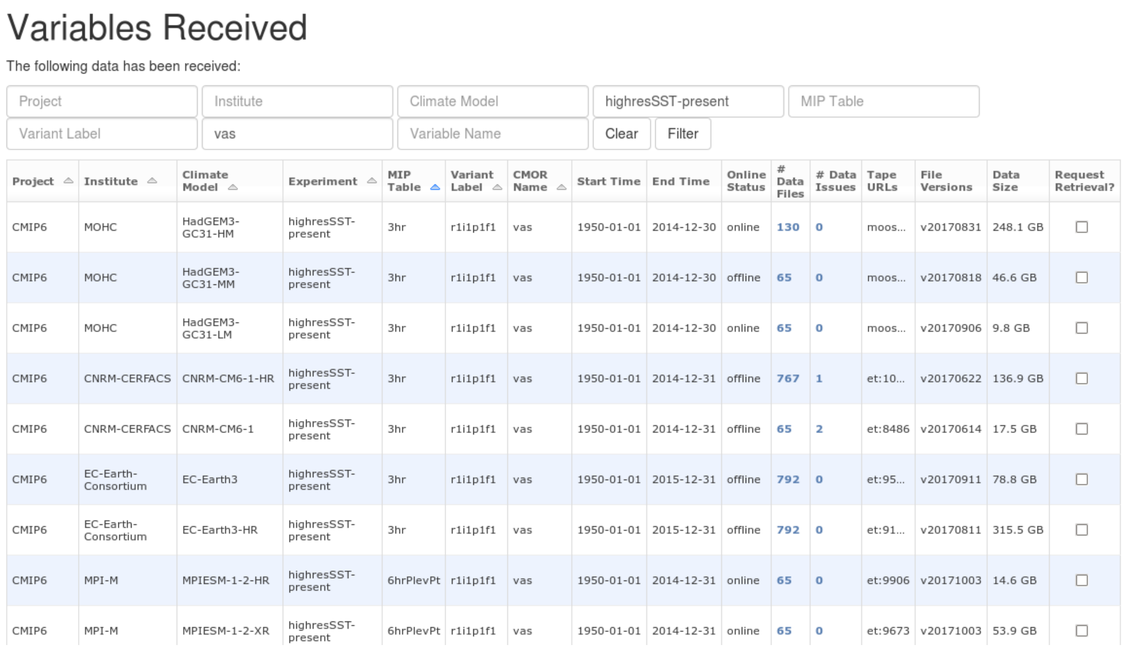
Data Management Tool

A Data Management Tool (DMT) has been developed and implemented on a dedicated server in the JASMIN managed cloud. The DMT consists of a database to store the metadata from each validated file and a web interface to allow data providers and data users to query the received data. The interface and database have been developed by PRIMAVERA funded staff at the Met Office and CEDA. It uses the Django web framework, Python and a PostgreSQL database.

Scientists use the DMT’s web interface to query what data is available. If the required data is only on tape then they can use the web interface to request that it is restored to disk. Figure 2 shows one of the DMT’s queries available to scientists; the data that has been received so far can be seen, along with its location and the facility to request that it is restored to disk. In this query the northward component of the near-surface wind speed (variable name: vas) in the highresSST-present (atmosphere only) experiment has been requested.

Recent Progress

- 1,254,287 output netCDF files have been uploaded to JASMIN.
- These contain 587 TB of climate data.
- 104 users are registered to access the data.
- 49 users have accounts on the Data Management Tool and have been actively uploading data or restoring data from tape to disk.



Variables Received

The following data has been received:

Project | Institute | Climate Model | highresSST-present | MIP Table

Variant Label | vas | Variable Name | Clear | Filter

Project	Institute	Climate Model	Experiment	MIP Table	Variant Label	CMOR Name	Start Time	End Time	Online Status	# Data Files	# Data Issues	Tape URLs	File Versions	Data Size	Request Retrieval?
CMIP6	MOHC	HadGEM3-GC31-HM	highresSST-present	3hr	r1i1p1f1	vas	1950-01-01	2014-12-30	online	130	0	moos...	v20170831	248.1 GB	☐
CMIP6	MOHC	HadGEM3-GC31-MM	highresSST-present	3hr	r1i1p1f1	vas	1950-01-01	2014-12-30	offline	65	0	moos...	v20170818	46.6 GB	☐
CMIP6	MOHC	HadGEM3-GC31-LM	highresSST-present	3hr	r1i1p1f1	vas	1950-01-01	2014-12-30	online	65	0	moos...	v20170906	9.8 GB	☐
CMIP6	CNRM-CERFACS	CNRM-CM6-1-HR	highresSST-present	3hr	r1i1p1f1	vas	1950-01-01	2014-12-31	offline	767	1	et:10...	v20170622	136.9 GB	☐
CMIP6	CNRM-CERFACS	CNRM-CM6-1	highresSST-present	3hr	r1i1p1f1	vas	1950-01-01	2014-12-31	offline	65	2	et:8486	v20170614	17.5 GB	☐
CMIP6	EC-Earth-Consortium	EC-Earth3	highresSST-present	3hr	r1i1p1f1	vas	1950-01-01	2015-12-31	offline	792	0	et:95...	v20170911	78.8 GB	☐
CMIP6	EC-Earth-Consortium	EC-Earth3-HR	highresSST-present	3hr	r1i1p1f1	vas	1950-01-01	2015-12-31	offline	792	0	et:91...	v20170811	315.5 GB	☐
CMIP6	MPI-M	MPIESM-1-2-HR	highresSST-present	6hrPlevPt	r1i1p1f1	vas	1950-01-01	2014-12-31	online	65	0	et:9906	v20171003	14.6 GB	☐
CMIP6	MPI-M	MPIESM-1-2-XR	highresSST-present	6hrPlevPt	r1i1p1f1	vas	1950-01-01	2014-12-31	online	65	0	et:9673	v20171003	53.9 GB	☐

Figure 2. The Data Management Tool’s web interface.

Data Analysis at JASMIN

The combination of the fast storage, interactive analysis servers and the LOTUS compute cluster allows PRIMAVERA scientists to bring their analysis to the data. There is no longer a need for scientists to download a copy of the data to their home institutes. However, because all of the data cannot be held on disk at once, the workflow shown in Figure 3 is necessary. This workflow is not as convenient as having all of the data set constantly online, but this is not possible because of the size of PRIMAVERA’s high-resolution data sets.

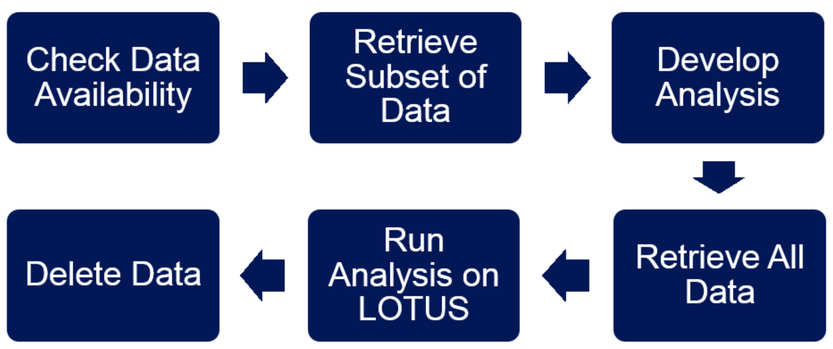


Figure 3. The workflow for analysing PRIMAVERA data.

Future Developments

- Publication of the data to the global community through the Earth System Grid Federation.
- Development of web based tools to allow additional users to examine and generate statistics from our multi-model ensemble of data. This is a further extension of the concept of users bringing their analysis to the data.
- Upload and analysis of additional ensemble members for some climate models.

Conclusions

JASMIN is the ideal environment to host the data storage and analysis facilities for the PRIMAVERA project because:

- its fast connections to the Internet allow the data to be rapidly brought together in one location;
- users can bring their analysis to the data, rather than having to download their own copy of the data, which isn’t feasible in a data set of this size;
- the JASMIN cloud allows custom user interfaces to the data to be developed;
- the JASMIN staff have the expertise to get the most out of the facility and out of the PRIMAVERA data.

PRIMAVERA
www.primavera-h2020.eu

PRIMAVERA is a collaboration between 19 leading European research and technology organisations with complementary expertise in climate science, climate change modelling, and high-performance computing. The project is led by the Met Office and the University of Reading.

Discover more about PRIMAVERA

PRIMAVERA has received funding from the European Union’s Horizon 2020 Research & Innovation Programme under grant agreement no. 641727.